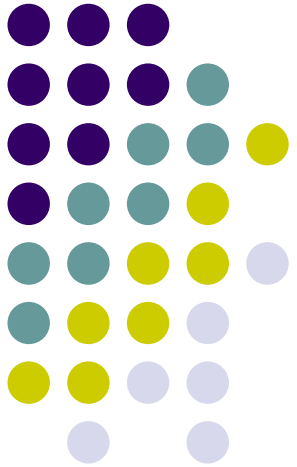
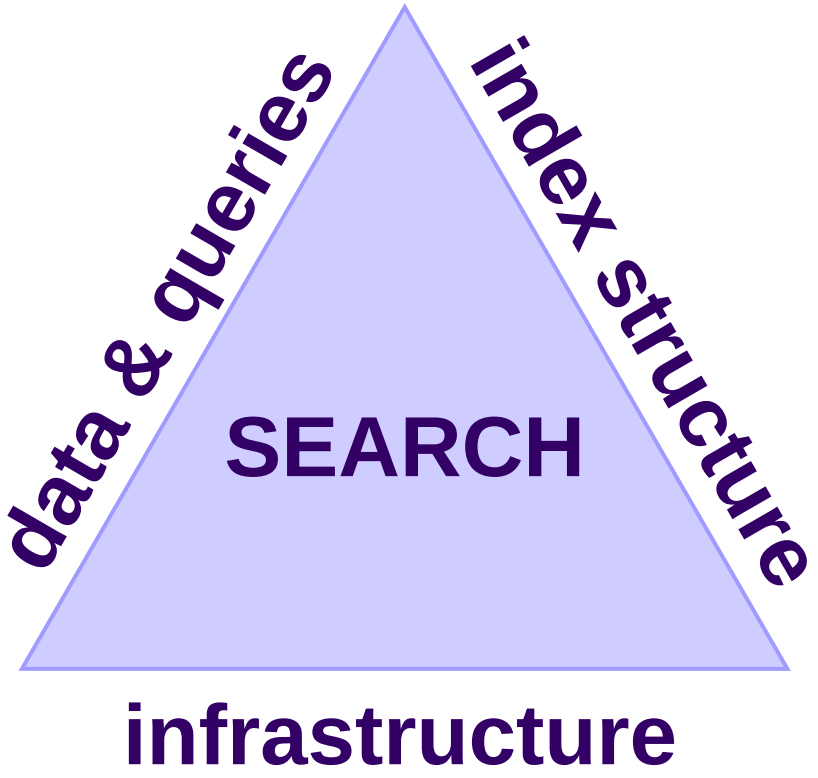
MUFIN Basics

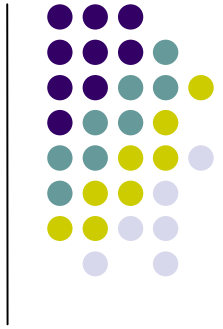
MUFIN team
Faculty of Informatics,
Masaryk University
Brno, Czech Republic
mufin@fi.muni.cz



Search problem



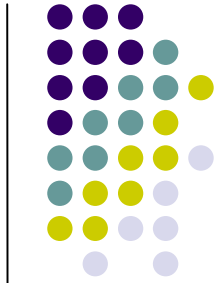
The thesis
(intellectual proposition)



- Search systems are more and more complex
- Future search system will be born on the divergence of:

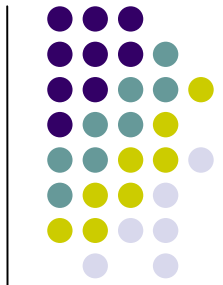
scale and **determinism**

Trends in Scalability of Search



- data volume - exponential growth
- number of users - increasing fast
- variety of data types - digital databases
- multi-queries - lingual, feature, modal

Trends in Determinism of Search



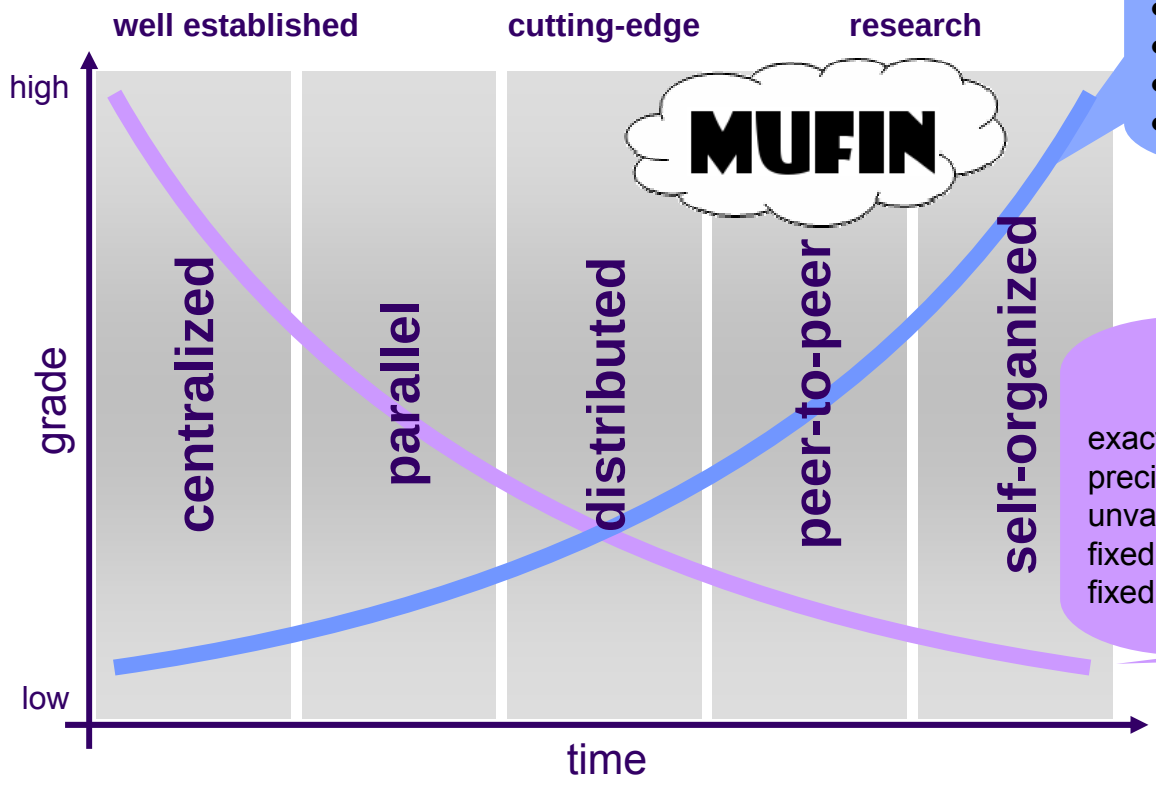
- Exact match
- Precise answer
- Unvaried answer

- Fixed query

- Dedicated hardware

- Similarity
- Approximate answer
- Satisfactory answer (advice, recommendation)
- Personalized, context aware, proximate
- Dynamic mapping, mobile devices, infrastructure services

Search systems



Scalability

- data volume – exponential grows
- number of users (queries) increase
- variety of data types - digitization
- multi-lingual (feature, modal) queries

Determinism

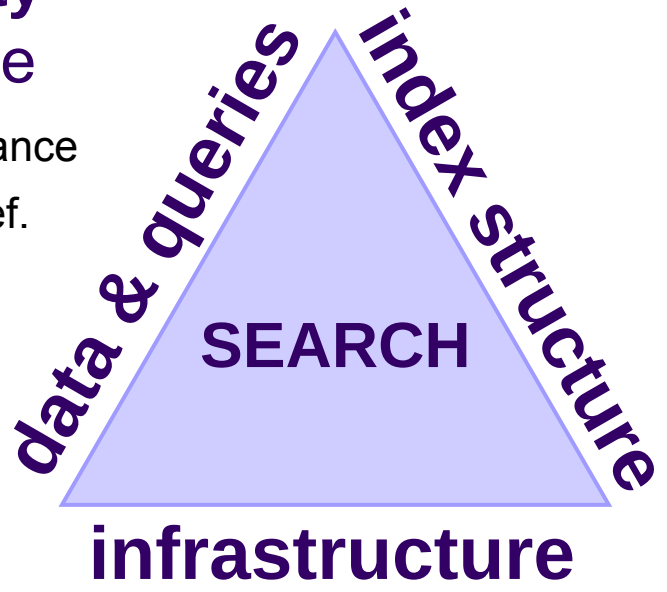
exact match	▸ similarity
precise	▸ approximate
unvaried answer	▸ good answer; advice
fixed query	▸ personalized; context aware
fixed infrastruct.	▸ dynamic mapping; mobile

The MUFIN Approach

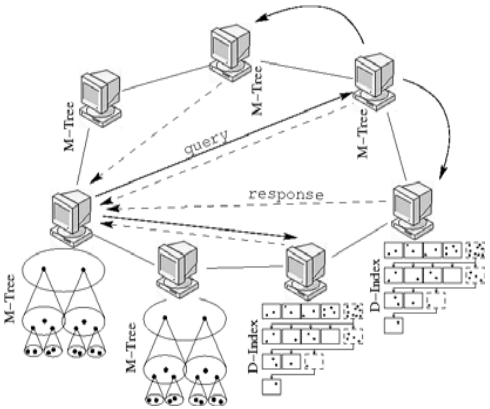
MUFIN: MUlti-Feature Indexing Network

Extensibility
metric space

Edit distance
Jaccard's coef.
Hausdforff distance
Minkowski distance
Mahalanobis distance
etc.

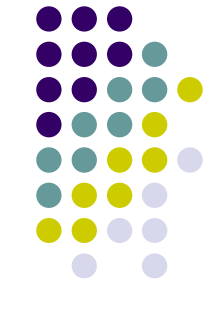


Scalability
P2P structure



Cloud computing
infrastructure as a service

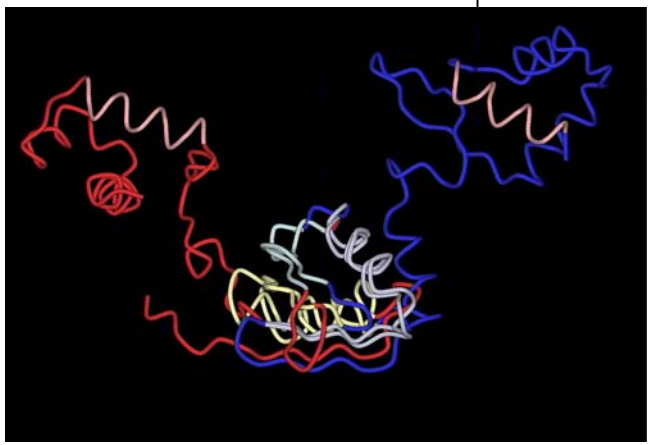
EXTENSIBILITY
Metric Space: Abstraction of Similarity



- Metric space: $\mathcal{M} = (\mathcal{D}, d)$
  - $\mathcal{D}$ – domain
  - distance function $d(x,y)$

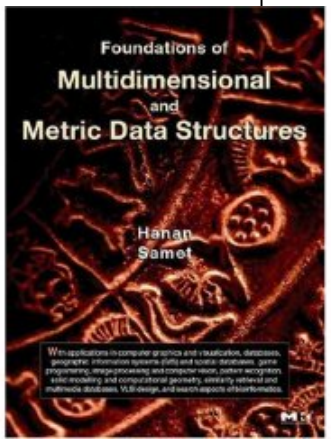
    $\forall x,y,z \in \mathcal{D}$
    - $d(x,y) > 0$ - *non-negativity*
    - $d(x,y) = 0 \Leftrightarrow x = y$ - *identity*
    - $d(x,y) = d(y,x)$ - *symmetry*
    - $d(x,y) \leq d(x,z) + d(z,y)$ - *triangle inequality*

Why Can the Metric Approach be Useful

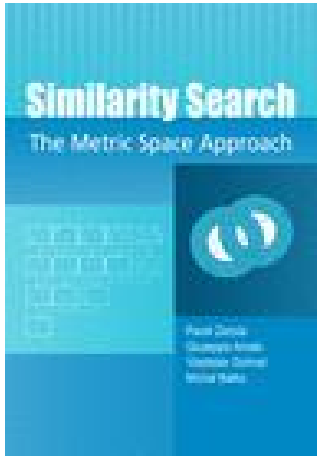


- Many application areas:
  - biology, security
  - audio-visual, geo. search
  - software copy detection
  - data cleaning, integration,
  - etc.
- Query by example paradigm
  - one query image contains a lot of information
  - one image is worth 1000 words
  - advantage for mobile devices – min. click

Metric Search Grows in Popularity



Hanan Samet
Foundation of Multidimensional and Metric Data Structures
Morgan Kaufmann, 2006



P. Zezula, G. Amato, V. Dohnal, and M. Batko
Similarity Search: The Metric Space Approach
Springer, 2006

Examples of Distance Functions

- L_p Minkowski distance of order p
  - L_1 – city-block distance
  - L_2 – Euclidean distance
  - L_∞ – infinity

$$L_1(x,y) = \sum_{i=1}^{n} | x_i - y_i |$$

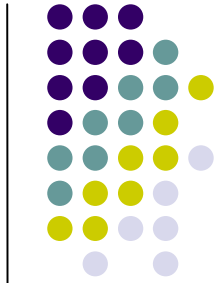
$$L_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$L_\infty(x,y) = \max_{i=1}^{n} |x_i - y_i|$$

- edit distance (for strings)
  - minimal number of insertions, deletions and substitutions
  - d('application', 'applet') = 6
- Jaccard's coefficient (for sets A,B)

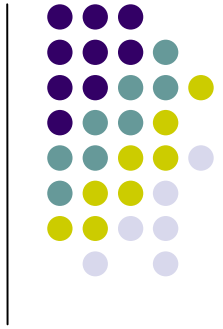
$$d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Examples of Distance Functions



- Mahalanobis distance
  - for vectors with correlated dimensions
- Hausdorff distance
  - for sets with elements related by another distance
- Earth movers distance
  - primarily for histograms (sets of weighted features)
- and many others

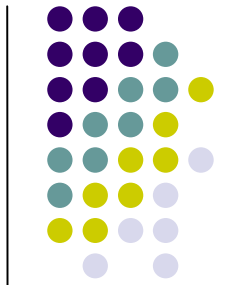
Image MUFIN overlay



- A demo on Cophir 50 M dataset (280 dim vectors)

- Five combined MPEG7 global descriptor:
  - Color Structure, max. dist.: 40, weight: 3
  - Color Layout, max. dist.: 300, weight: 2
  - Scalable Color, max. dist.: 3000 weight: 2
  - Edge Histogram, max. dist.: 68, weight: 4
  - Homogeneous Texture, max. dist.: 25, weight: 0.5

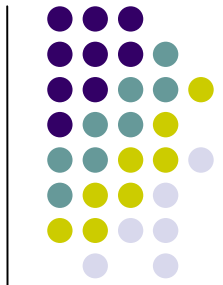
Face search



- Face search demo – 6k images with people
  - face detection – 10k detected faces
  - face description – 64 dimensional vectors
  - face comparison - *advanced face* des. MPEG7

- Based on a publicly available software

SCALABILITY
Structured P2P networks



- Objectives
  - To scale into contemporary audio-visual data volume and query execution throughput, i.e.:
    - billions of objects
    - online response time
    - hundreds of queries per sec.
- A peer
  - Contains metric objects, can issue/answer queries, and knows few other peers

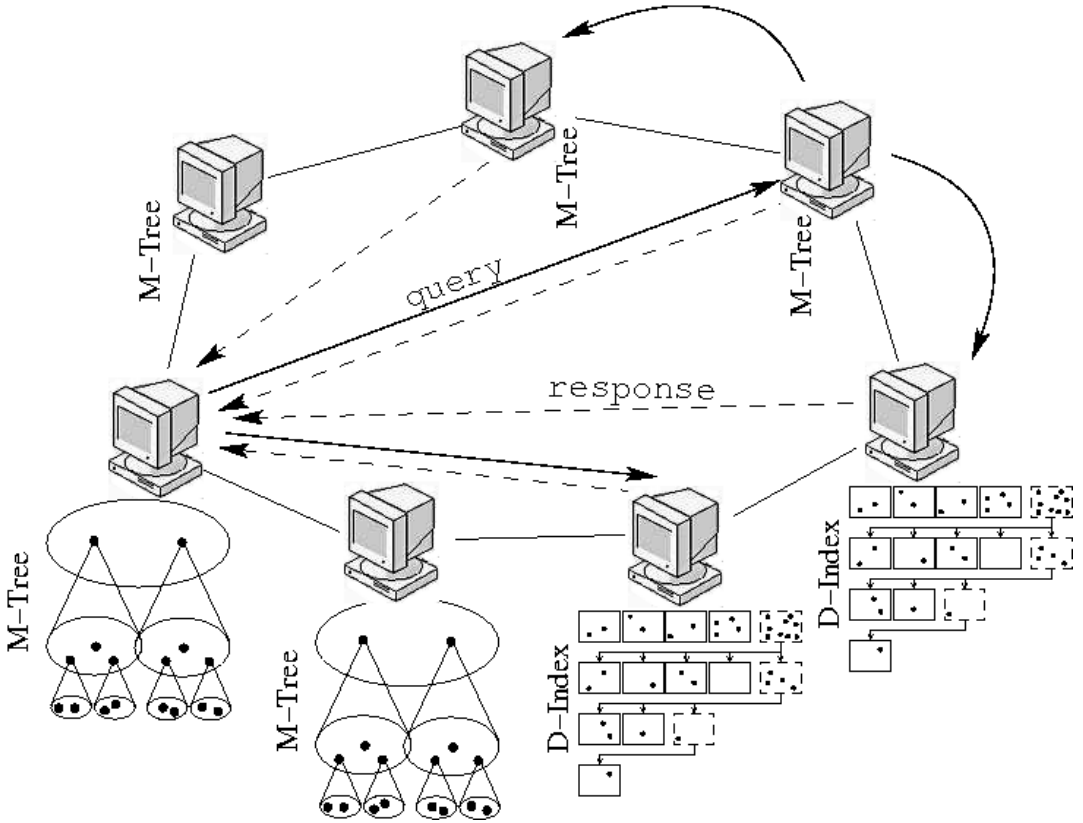
Why structured P2P in MUFIN

- Structured P2P network employ a globally considered protocol to ensure that any peer can efficiently route a search to some peer that has the desired data

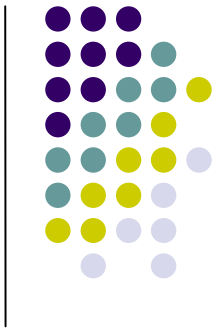
- Structured P2P networks are used in MUFIN for:
  - no bottleneck, no central component
  - multiple access points to the networks
  - distribution of workload – parallel query execution
  - dynamic structure of peers – (controlled) resilience, join, leave
  - mechanisms for fault tolerance, replication and load balancing

P2P Architecture of MUFIN

- Native metric techniques: **GHT*, VPT***
- Transformation techniques: **MCAN, M-Chord (Skip-Graphs, Kademlia, etc.)**

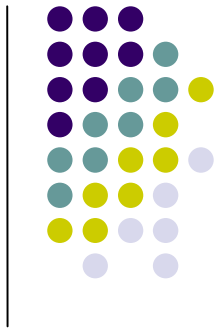


P2P Architecture of MUFIN



- Peers are not necessarily computers
- A peer size determines a lower-bound on the query response time
- Peer's data can be searched by:
  - Filtering
  - M-tree
  - D-index
  - I-distance
  - Etc.

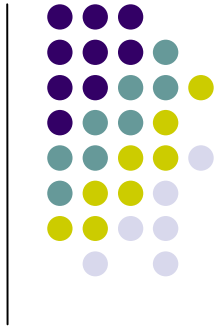
Scalability test



- 1M: 50 peers – memory based
- 10M: 500 peers – memory based
- 50M: 2000 peers – disk based

- Effectiveness improves with data volume
- Efficiency
  - lower-bounded by the peer size (20k, 20k, 25k)
  - does not change significantly

Infrastructure as a Service

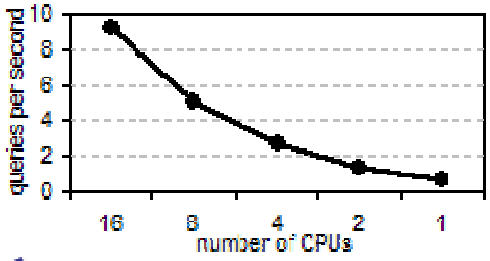


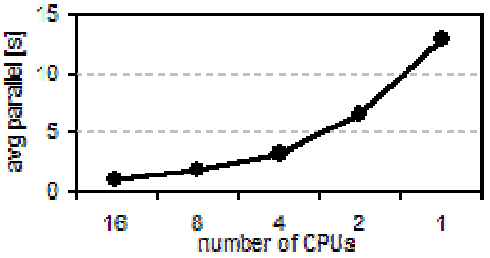
- Why:
  - Performance tuning
    - Query response time
    - Query execution throughput
  - Performance adjustment
    - Different performance requirements (day – night, weekend – working days)
  - Experimental trials
    - Test an application
    - Purchase a new hardware
  - Availability - reliability

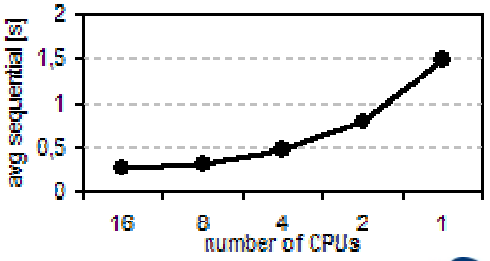
MUFIN Hardware Mapping

- 10M network, 500 peers, memory-based
- Batch of 250 queries started from 10 peers

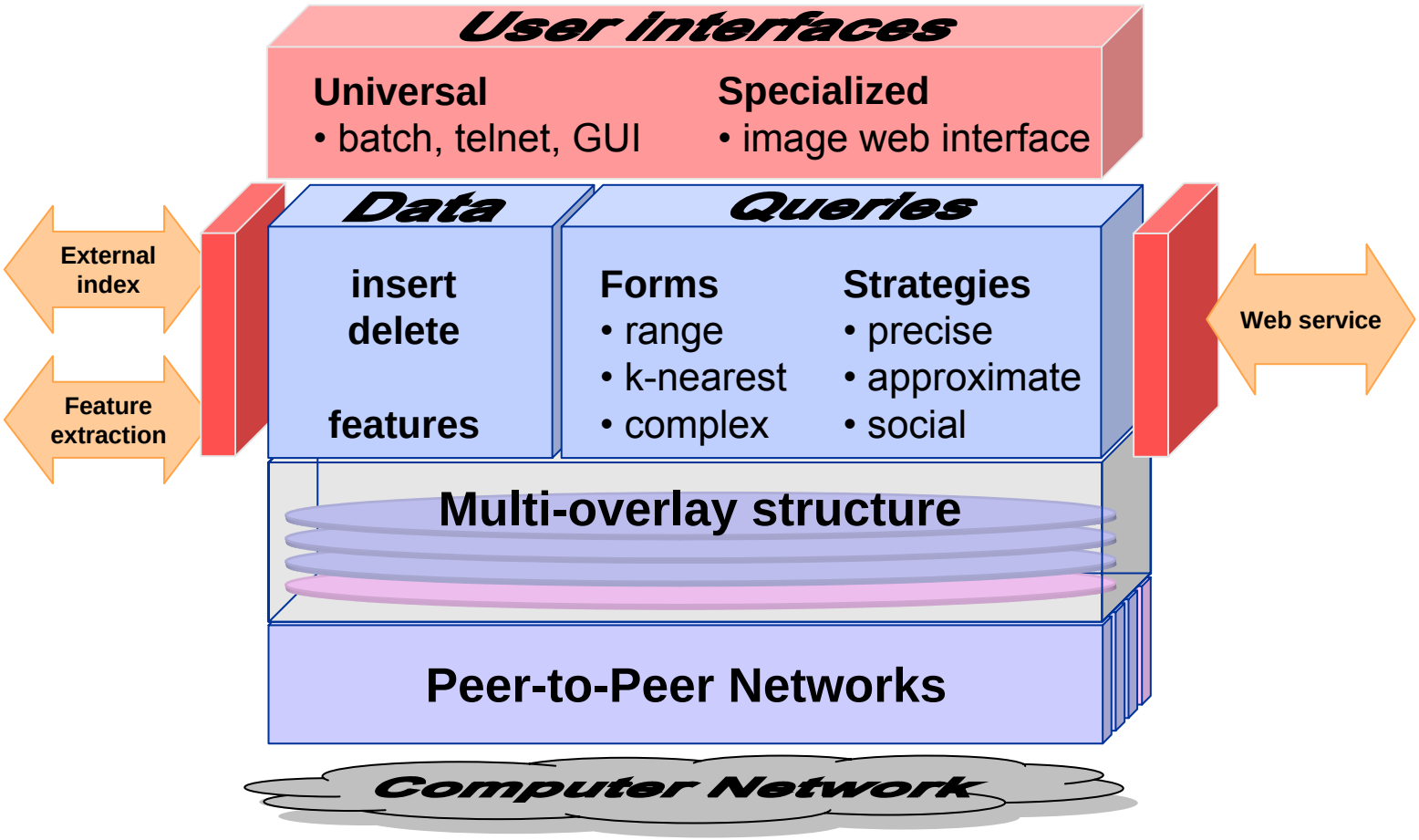
CPUs	Parallel from 10 peers					Sequential from 1 peer			
	total [s]	queries/s	single query [ms]			total (s)	single query [ms]		
			avg	min	max		avg	min	max
16	27	9,26	958	184	2691	67	259	183	1605
8	49	5,10	1787	181	5736	87	324	170	1806
4	94	2,66	3265	165	10355	122	468	162	1847
2	186	1,34	6654	165	24483	203	780	168	2320
1	380	0,66	12810	169	69692	379	1472	169	3248



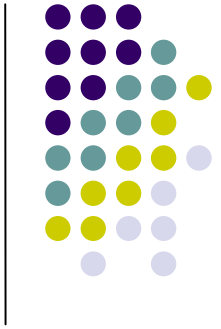




MUFIN Overview



MUFIN plugin



- News web-sites contain images
  - CNN, BBC, SEZNAM, iDNES
- Photography collection of US National Parks
  - TERRA GALLERIA
- Image text search
- Google, Yahoo, Yandex, Ask, Seznam, Rajče, exalead

Use of MUFIN in SAPIR Demos

- Caching – to locate cashed queries
- Text+Image – to perform content similarity
- Video search - to perform content similarity
- Mobile interface - to perform content similarity

- Some statistics
- Permanent demo: coming soon